Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 11869–11873, December 1994
Cell Biology

Isolation of unknown genes from human bone marrow by differential screening and single-pass cDNA sequence determination

SALLY L. ORR*†, TIM P. HUGHES‡, CHARLES L. SAWYERS‡, ROBERTA M. KATO‡, SHIRLEY G. QUAN‡, SUZANNE P. WILLIAMS§, OWEN N. WITTE‡, AND LEROY HOOD¶||

*Division of Biology, California Institute of Technology, Pasadena, CA 91125; ‡Howard Hughes Medical Institute, University of California, Los Angeles, CA 90026; §Central Research Division, Pfizer Inc., Groton, CT 06340; and ¶Department of Molecular Biotechnology, University of Washington School of Medicine, Seattle, WA 98195

Contributed by Leroy Hood, August 5, 1994

ABSTRACT **A cDNA sequencing project was initiated to characterize gene expression in human bone marrow and develop strategies to isolate novel genes. Forty-eight random cDNAs from total human bone marrow were subjected to single-pass DNA sequence analysis to determine a limited complexity of mRNAs expressed in the bone marrow. Overall, 8 cDNAs (17%) showed no similarity to known sequences. Information from DNA sequence analysis was used to develop a differential prescreen to subtract unwanted cDNAs and to enrich for unknown cDNAs. Forty-eight cDNAs that were negative with a complex probe were subject to single-pass DNA sequence determination. Of these prescreened cDNAs, the number of unknown sequences increased to 23 (48%). Unknown cDNAs were also characterized by RNA expression analysis using 25 different human leukemic cell lines. Of 13 unknown cDNAs tested, 10 were expressed in all cell types tested and 3 revealed a hematopoietic lineage-restricted expression pattern. Interestingly, while a total of only 96 bone marrow cDNAs were sequenced, 31 of these cDNAs represent sequences from unknown genes and 12 showed significant similarities to sequences in the data bases. One cDNA revealed a significant similarity to a serine/threonine-protein kinase at the amino acid level (56% identity for 123 amino acids) and may represent a previously unknown kinase. Differential screening techniques coupled with single-pass cDNA sequence analysis may prove to be a powerful and simple technique to examine developmental gene expression.**

Large-scale cDNA sequencing is an important part of the Human Genome Project. The sequence identification of all expressed cDNAs will benefit a wide variety of research, including physical mapping, gene structure analysis, and the identification of the genetic basis for many diseases (1, 2). Of the approximately 100,000 human genes, an estimated 29,000 human sequences are present in the data bases. To isolate every human gene, it will be necessary to construct cDNA libraries from a variety of tissues and developmental states. In addition, the use of sophisticated cDNA library construction and library screening techniques will also be necessary to decrease cDNA sequence redundancy and increase the probability of isolating rare cell-specific cDNAs. It is likely that each tissue source will pose unique problems before all human cDNAs are isolated. Thus combinations of cellular enrichment techniques, subtracted or normalized libraries, and differential screening will be necessary to isolate rare human mRNAs. Recent human cDNA sequencing projects have been successful at isolating genes expressed in the brain (1, 3), testis (4), T cells (5), and retinal epithelium (6).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Bone marrow (BM) is a complex tissue containing different cell types and cells at different stages of differentiation. The major cellular groups are hematopoietic cells and stromal cells. In addition, BM contains rare hematopoietic stem cells (HSC), the precursors of erythroid, myeloid, and lymphoid cell populations. Only a fraction of the genes which control the growth and differentiation of hematopoietic cells within human BM have been identified. Moreover, the complete DNA sequence information, chromosomal location, and cell-type expression patterns of BM genes will be a valuable resource for understanding the molecular basis of BM-derived disorders.

This project was initiated to determine the complexity of total BM sequences and determine the efficacy of differential screens to remove redundant sequences and thereby enrich for rare lineage-specific mRNAs. Previous large-scale single-pass cDNA sequence analysis has resulted in the generation of over 20,000 expressed sequence tags (ESTs), the majority of which are from human cDNA libraries. "EST" has become an unfortunate nomenclature, since the presence of these sequences in cDNA libraries does not guarantee that they are derived from true *expressed* transcripts. Contaminants may arise from genomic DNA, incomplete splicing, or low-level contamination from various microorganisms. To circumvent these problems we have tested coupling single-pass DNA sequence determination to rapid expression studies on Northern blots. Unique cDNA sequences with restricted expression patterns will be subject to further structure/function analysis. We show that with as few as four runs on an automated DNA sequencer, or 96 cDNAs sequenced, 31 of these may correspond to previously unknown genes, and one of these is likely to encode a serine/threonine kinase.

MATERIALS AND METHODS

Construction and Isolation of Human BM cDNAs. Human BM was obtained from a healthy male and a healthy female donor. Total RNA was isolated (7) and poly(A)$^+$ RNA was purified by two passes of oligo(dT) chromatography. One-half microgram of poly(A)$^+$ RNA was used to synthesize cDNA with random hexamers as described by the TimeSaver cDNA synthesis kit (Pharmacia). Random hexamers (50 ng) were used for 2 hr of cDNA synthesis and oligo(dT) was added during the last hour. Complementary DNA was ligated into M13mp18 and stored at −70°C. Transformations were performed with 0.5 μl of the 35-μl ligation mixture, using

Abbreviations: BM, bone marrow; PSBM, prescreened bone marrow; EST, expressed sequence tag; GRAIL, Gene Recognition and Analysis Internet Link.

†Present address: Central Research Division, Pfizer Inc., Groton, CT 06340.
||To whom reprint requests should be addressed.

supercompetent XL1-Blue cells (Stratagene) and titrated to a density of 500 plaques per plate. Individual plaques were picked and tested for inserts by PCR using M13 (*lacZ*) primers (BV40, 5'-ACGTTGTAAAACGACGGCCAGT-3'; BV41, 5'-GAAACAGCTATGACCATGATTA-3').

Differential Hybridization. For differential screens, 48 BM cDNAs were sequenced and subject to DNA sequence analysis. Twenty-seven of these identified BM cDNAs were chosen to represent known human sequences and unwanted cDNAs such as globin. Two probes were prepared and mixed in the final hybridization. The first probe consisted of gel-purified PCR products of the 27 BM cDNAs with known sequences, pooled and labeled by random priming. The second probe was synthesized by using 2.4 μg of HeLa cell poly(A)$^+$ RNA and avian myeloblastosis virus reverse transcriptase (Molecular Genetics) to a specific activity of 1.5×10^7 cpm/μg, and the entire cDNA reaction mixture was added to the hybridization mixture with the first BM probe. Differential hybridizations, of 10^4 M13 phage plated to a density of $\approx$500 plaque-forming units per plate, were performed in 5× SSPE/5× Denhardt's solution containing salmon sperm DNA at 100 μg/ml, 10% dextran sulfate, and 0.1% SDS for 16 hr at 58°C. Filters were washed once in 2× SSC/0.1% SDS at 25°C and three times in 0.2× SSC/0.1% SDS at 65°C. (1× SSPE = 150 mM NaCl/10 mM sodium phosphate, pH 7.4/1 mM EDTA; 1× Denhardt's solution = 0.02% bovine serum albumin/0.02% Ficoll/0.02% polyvinylpyrrolidone; 1× SSC = 150 mM NaCl/15 mM sodium citrate, pH 7.0.) Plates contained 5–10% background nonrecombinants.

RNA Gel Blot Hybridization. Total RNA was isolated from cell lines (7) and 20 μg was electrophoresed on formaldehyde/1% agarose gels. RNA was transferred to nitrocellulose and probed with gel-purified PCR products labeled by random priming. Probes were added to hybridization mixtures containing 50% deionized formamide, 5× SSC, 1× Denhardt's solution, 50 mM sodium phosphate at pH 7.0, and salmon sperm DNA at 0.25 mg/ml and hybridized for 12–16 hr at 42°C. Blots were washed in 2× SSC/0.1% SDS at room temperature for 30 min and in 2× SSC/0.1% SDS at 60°C for 30–60 min. Cell lines were kind gifts from the following sources: KG1 (8, 9) and M07E, provided by D. Golde (Sloan–Kettering); AF10, provided by A. Saxon [University of California, Los Angeles (UCLA)]; Jurkat (10), provided by W. Clark (UCLA); CCRF-CEM (11), provided by C. Uittenbogaart (UCLA); PLB 985 (12), provided by S. Smale (UCLA); and BV173 (13) and IMR 5, provided by A. Van Herle (UCLA). Also used were ALL-1 (14) and HEL (15).

cDNA Sequence Determination and Data Base Analysis. For single-run DNA sequence determination, M13 DNA was prepared using a PEG miniprep procedure starting with 1.0 ml of infected culture. Sequence determination was performed by using *Taq* DNA polymerase, M13 fluorescent primers (Applied Biosystems), and the Applied Biosystems catalyst. Reactions were run on the Applied Biosystems 373 automated DNA sequencer. Approximately 400 bases were compared with nonredundant European Molecular Biology Laboratory and GenBank data bases (version 73), using BLAST (16). Sequences were also translated and used to search the protein data bases (Protein Identification Resource), using the BLASTX sequence analysis program. DNA sequences that were determined to be unknowns from the BLAST analysis were subjected to further analysis using the FASTA program (17). In addition, all unknown sequences were sent by electronic mail to GRAIL (Gene Recognition and Analysis Internet Link) (18). Predicted exons were translated and further protein analysis was performed by using the FASTA program.

RESULTS

Partial DNA Sequence Analysis of BM cDNAs. Random shotgun DNA sequencing of cDNAs is straightforward and is based upon the assumption that if the sequence was prepared from mRNA, it represents an expressed gene. The drawbacks of limited shotgun sequencing are the loss of expression information, possible sequencing of noncoding regions, and the generation of sequence data from different regions of the same transcript. We are beginning a cDNA sequencing project using cDNAs isolated from human BM. Since we are interested in isolating rare mRNAs that are expressed by the developing hematopoietic system, we are evaluating combinations of cDNA library prescreening coupled with rapid methods for expression analysis. Forty-eight BM cDNAs were subject to single-pass DNA sequence determination. Eight (17%) of these represented cDNA sequences not found in searching the GenBank data bases, using the BLAST sequence analysis program (Table 1). The major mRNA expressed in total BM was determined to be globin. Globin was found to be present in 23% of the cDNAs sequenced. Globin mRNA has been measured at 3% of murine reticulocyte RNA, and up to 22,600 copies of globin mRNA have been found per cell (19). Our result may be higher than expected given a limited cDNA sample size. Other known mRNAs observed in the random library include mitochondrial genes, ribosomal sequences, and repeats. Additional abundant sequences were observed that may be unique to BM-derived cDNAs. In addition to globin, these abundant BM mRNAs include eosinophil basic protein and immunoglobulins.

Initial cDNA sequencing and analysis of the total BM library revealed problems associated with large-scale sequencing, such as the repetitive sequencing of common "housekeeping" and other abundant mRNAs. To eliminate these abundant unwanted sequences and increase the number of rare, cell-specific, and possibly unknown sequences, a differential screening strategy was tested. A complex probe consisting of two cDNA mixtures was designed to eliminate abundant sequences. This probe contained labeled cDNAs from the known BM sequences (Table 1) and labeled cDNA

Table 1. Comparison of BM and prescreened BM (PSBM) DNA sequence data to nucleic acid data bases

	BM cDNAs		PSBM cDNAs	
Type of sequence match	Number	%*	Number	%*
Unknown sequence/no match	8	17	23	48
Human sequences	8	17	9	19
Immunoglobulin	2		3	
Eosinophil basic protein	2		0	
Others	4		6	
Nonhuman sequences†	2	4	2	4
Unwanted sequences	30	63	14	29
Globin	11		3	
Mitochondrial	6		0	
Ribosomal	3		3	
Actin/tubulin	3		0	
Alu repeats	2		5	
Others‡	5		3	
Total	48		48	

BM and PSBM cDNA sequence data, generated from one sequence run, were compared with sequences in a nonredundant DNA data base by using the BLASTN (nucleic acid) and BLASTX (protein) analysis programs.

*Percent of 48.

†Nonhuman data base matches indicate similarities to nonhuman DNA sequences, the lowest being 69%/178 bp and the highest 87%/40 amino acids.

‡Other sequence data include pieces of vector sequences (nonrecombinants) and poly(A) runs.

derived from HeLa cells. The HeLa probe was to identify abundant sequences present in rapidly growing human cell lines that might have been missed in our initial BM cDNA sequencing. Since complex probes generated from total mRNA populations cannot identify rare sequences, this screen should label abundant sequences (≥0.2%). The two probes were mixed and hybridized to the total BM library. Approximately 53% of the M13 plaques appeared as slight to no hybridization. When more conservative criteria were used, 10% of the M13 plaques were negative. Following the differential screen, plaques that failed to hybridize with the complex probe were isolated. Forty-eight of these PSBM cDNAs were subject to limited DNA sequence determination (Table 1). As expected, the prescreen increased the unknown cDNA sequences from 17% to 48%. The percentage of cDNAs with data base matches to known human sequences and significant matches to nonhuman genes were unchanged after the differential prescreen. The number of unwanted sequences decreased from 63% to 29%. Interestingly, globin sequences were not completely removed. The number of sequences containing *Alu* repeats were also slightly increased after the prescreen. This may reflect the increased sequence complexity derived from unspliced poly(A)$^+$ nuclear sequences.

For data analysis, cDNA sequences were first examined for similarities to known genes by using the BLAST program for nucleic acids. In addition, all six reading frames were examined for similarities to proteins by using the BLASTX program. To examine the accuracy of one-pass DNA sequence determination, six cDNAs encoding known human sequences were aligned, using both BLAST and FASTA sequence analysis programs. The average BLAST alignment was 95 ± 5% in 292 bases and the FASTA average was 92 ± 6% in 356 bases. The alignment capabilities of the FASTA program were also used to test whether the sequences of unknown cDNAs were redundant. Cross-sequence analysis revealed that the highest overlap between unknown cDNAs was 39% in 621 bases. Therefore it is unlikely that an abundant unknown transcript was sequenced multiple times, unless the cDNAs prepresent nonoverlapping sections of abundant unknown clones. The sequences of 15 random unknown cDNAs were also analyzed to determine whether they contained open reading frames. All of these cDNAs had open reading frames, with an average length of 206 ± 60 bases.

Once BLAST analysis failed to reveal significant similarities between cDNA sequences and sequences in the data bases, the unknown sequences were then subject to further analysis, including GRAIL and FASTA. Twenty-two sequences from unknown cDNAs were sent to the GRAIL system for putative exon prediction. The GRAIL system uses a multiple sensor neural network to identify protein-coding regions of at least 100 bases and has proven to accurately predict exons 89% of the time (18). Eight of the 22 cDNAs examined by GRAIL contained potential exons (Table 2). One explanation why many of the sequences failed to contain predicted protein-coding regions may be that these sequences are derived from noncoding regions. These contributions could be from 3′ or 5′ untranslated regions or from incompletely spliced poly(A)$^+$ RNAs, which are frequently observed in BM cDNA libraries. The quality of the exon predictions according to GRAIL varied equally from "excellent" to "good." Since protein data base searches are more sensitive than nucleic acid searches, the translated protein sequences of predicted exons following GRAIL analysis were examined for similarities to known proteins by using the FASTA program. While this exon analysis regime failed to find a significant similarity to a known sequence, one cDNA exhibited short and multiple similarities to Na/K-ATPases (data not shown). Overall, the most important criterion to prioritize unknown cDNAs for further analysis is a significant BLASTX (protein) similarity. However, the presence of predicted exons and multiple weak similarities to known genes can also be used to prioritize unknown cDNAs for further RNA expression analysis employing Northern blots.

Table 2. Exon prediction of unknown BM cDNAs by using GRAIL

Clone	Strand	Probability*	Frame	Quality†	ORF
BM1	F	0.76	3	Excellent	318–525
PSBM11	F	0.76	1	Excellent	136–214
PSBM19	F	0.85	3	Excellent	1–286
PSBM23	R	0.85	2	Excellent	1–170
PSBM27	R	0.82	2	Good	161–400
PSBM38	R	0.77	3	Good	261–555
PSBM41	R	0.71	3	Good	432–480
PSBM46	F	1.00	1	Good	1–241
PSBM46	R	0.79	2	Good	1–281

GRAIL uses a multiple-sensor neural network to identify coding exons of at least 100 bp in human DNA. Results shown are from 22 unknown cDNA sequences submitted to GRAIL by electronic mail. F, forward strand; R, reverse strand; ORF, open reading frame (base pair numbers).

*Probability score greater than 0.5 identifies a region with protein-encoding potential.

†Quality equals the GRAIL systems interpretation of raw data, taking into account strand, probability, frame, and ORF data.

RNA Expression Analysis of Unknown cDNAs. Following DNA sequence analysis, unknown cDNAs were tested for expression against a panel of human leukemic cell lines. These cell lines were chosen to represent a variety of different hematopoietic cell types and differentiation states, including multipotent cell lines. First, PCR products of cDNAs were labeled and tested on RNA blots, using a limited number of cell lines. This preliminary screen was used to quickly assay whether the expression was restricted to a cell type or was ubiquitously expressed among all cell lines tested. Following this preliminary screen, the strategy was to increase the number of cell lines tested and probe with cDNAs that exhibited restricted or no expression. Of 13 unknown cDNAs tested, 10 were expressed in all cell lines tested and 3 revealed a restricted expression pattern (Fig. 1). In addition, one cDNA, PSBM6, exhibited two transcripts on RNA blots. Further analysis of the 3 cDNAs with restricted expression revealed major but not exclusive expression in early myeloid cell lines. This may reflect the predominance of myeloid cells over erythroid and lymphoid cells in nucleated BM populations (20). One caveat concerning using mRNA from continuous cell lines is that aberrant gene expression may limit their value to assign gene expression to defined cell types. As a consequence, it becomes necessary to include multiple cell lines representing the same lineage or state of differentiation on RNA blots. However, the ease in growing large quantities of these cells to isolate mRNA to use in RNA blots makes these cell lines ideal for an initial prescreen. To characterize lineage-specific expression further, it will be necessary to use panels of normal human BM and peripheral blood cells and to enrich for specific cell types either by sorting or by selective culture conditions. In addition, further analysis will include complete DNA sequence determination and chromosomal mapping.

DNA Sequence Analysis and the Identification of a Protein Kinase. Single-pass DNA sequence analysis of 96 BM cDNAs chosen by random selection and differential screening revealed known human genes, known genes originally isolated in other species, and unknown cDNA sequences. cDNAs with significant similarities to nonhuman genes may represent the human equivalents of these genes. As an example, the translation of BM9 revealed 126/141 (89%) amino acid similarity to mouse Tum p198, a transplantation antigen

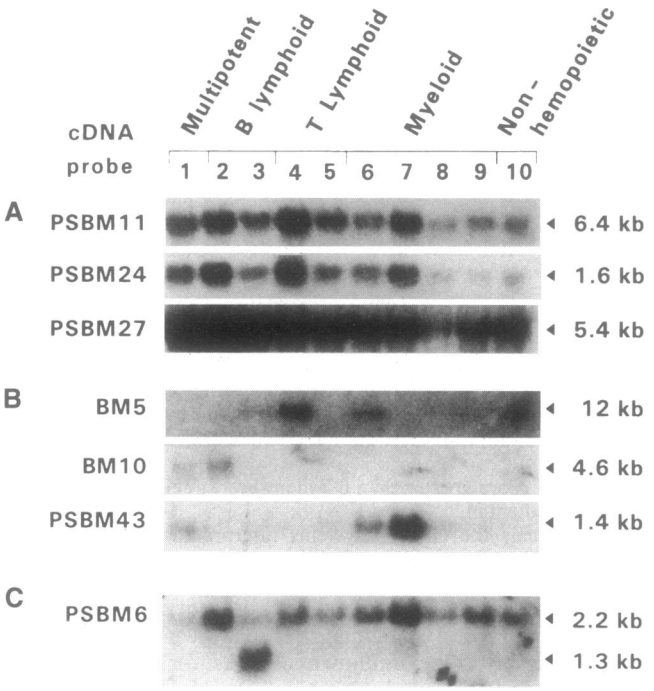

FIG. 1. Northern blot analysis of individual BM cDNAs. Leukemic cell lines used were as follows: multipotent, BV173 (lane 1); pre-B lymphoid, ALL-1 (lane 2) and plasma cell AF10 (lane 3); T lymphoid, CEM (lane 4) and Jurkat (lane 5); myeloid, KG-1 (lane 6), megakaryocytic M07E (lane 7), myelomonocytic PLB 985 (lane 8), and HEL (lane 9); and nonhemopoietic, neuroblastic IMR-5 (lane 10). (*A*) Northern analysis of cDNAs that are generally expressed in leukemic cell lines. (*B*) Northern analysis of cDNAs expressed in a cell-lineage-restricted expression pattern. (*C*) Northern analysis of PSBM6, a probable new kinase, demonstrating two transcripts in the AF10 cell line.

expressed by the P815 mastocytoma (data not shown) (21). Of the unknown or unique cDNA sequences, 16 exhibited significant similarities to known genes in the data bases (Table 3). Interestingly, with an estimated >20,000 unknown cDNAs or EST sequences present in the data bases, only 1 of the 96 BM cDNA sequences exhibited significant similarities to an EST. This may reflect gene expression differences in the brain, where most of the ESTs have been derived, versus the BM. Moreover, since the majority of the ESTs were generated from libraries that used oligo(dT) as a primer, a large percentage of ESTs may correpond to 3′ untranslated sequences. We have chosen random primers to synthesize cDNA of moderate size to increase the possibility of generating sequences from exons and thus aid in the identification of novel genes.

Of the three cDNAs that exhibited lineage-restricted expression patterns by Northern analysis, only one revealed any significant similarity to a known gene. BM10, a 4.6-kb transcript expressed primarily by B lymphoblastoid cells, has significant similarity to a pig gene expressed during hepatic stress (Fig. 2).

Of the 16 cDNAs that exhibited significant similarities to known genes, one cDNA was of particular interest. PSBM6 was observed to be similar to protein kinase sequences on both the nucleic acid and protein level. In addition, multiple significant similarities to other kinases were present throughout the sequence. The highest similarity was a 56%/123 amino acid identity to human p78 (Fig. 2). This kinase is an unpublished data base entry described as a 78-kDa marker lost in a chemically induced transplantable carcinoma and related to serine/threonine-protein kinases. Therefore this cDNA probably encodes a nonreceptor kinase that may catalyze protein phosphorylation in response to second messengers.

SUMMARY AND CONCLUSIONS

cDNA Analysis. Recently large-scale cDNA sequencing projects have contributed thousands of short (>150- to 500-bp) cDNA sequences to the data bases (1, 3, 22). We have sequenced 98 cDNAs corresponding to four runs on the Applied Biosystems 373 DNA sequencer or 10 manual sequencing gels. This limited scale can be easily accomplished by most laboratories. Of these 98 cDNAs sequenced, 31 cDNA sequences corresponding to unknown genes were isolated. While the percentage of unknown cDNA sequences isolated will vary with different tissues and even different libraries of the same tissue, the identification of a previously unknown kinase demonstrates that shotgun cDNA sequencing in BM revealed novel cDNAs with significant sequence similarities to known genes at a frequency of 1%. Moreover, even a limited sequencing approach, when coupled to expression analysis, can reveal genes with restricted expres-

Table 3. Unknown BM cDNAs and PSBM cDNAs with significant similarities to known genes

cDNA clone	Known gene or product			Overlap length	% identity
	Accession	Description	Species		
BM1	YSCMTCG	Mitochondrial DNA	Yeast	39 nt	76
BM3	B39057	Amelogenin	Human	28 aa	60
BM9	M51528	ATP phosphotransferase	Rat	255 nt	81
BM10	PIGHEP1	Unidentified hepatic mRNA	Pig	67 nt	85
BM22	HUMACROD	Acrosin gene, 3′ end	Human	16 aa	75
BM26	M73998	CAP18 protein mRNA	Rabbit	178 nt	69
BM28	M31166	TNF-inducible mRNA	Human	24 aa	79
PSBM2	HSB37G022	EST partial cDNA sequence	Human	222 nt	96
PSBM3	SCMP48EGG	p48 eggshell protein gene	*S. mansoni*	17 aa	29
PSBM6	HUMP78A	Protein p78 mRNA	Human	123 aa	56
PSBM12	BOVRHOGDI	Dissociation inhibitor of rho	Bovine	302 nt	72
PSBM19	HSBBICP4A	Herpesvirus transcription control protein	Bovine	45 aa	28
PSBM28	M92295	γ1 and γ2 globin	Gorilla	52 nt	76
PSBM45	PFAMSAAAC	Major merozoite surface antigen	*Plasmodium*	64 nt	70
PSBM46	M78084	cDNA clone HHCP	Human	182 nt	63
PSBM48	RATASI	Amino acid starvation-induced protein	Rat	57 aa	59

BLASTN and BLASTX results of unknown BM cDNAs with similarities to known genes with a $P(N) \leq 0.01$. The $P(N)$ value represents the probability of a score occurring by chance, given the size of the sequence searched and the size of the data base. The yeast is *Saccharomyces cerevisiae*; TNF, tumor necrosis factor; *S. mansoni*, *Schistosoma mansoni*.

A

```
                140       150       160       170       180       190
BM10    TCATGCCCAGTTAACTTATTTACAATATTTAAGTTCTCTGCTTCTGCATTTGGTTGGTTT
                                    ::::: :::::::::::: :: ::: :::::
PIGH    ACGATTGTTACCACGCGCCATTGGTTGCGCAAGTTTTCTGCTTCTGCTTTGGGTGGGTTT
              1060      1070      1080      1090      1100      1110

                 200       210       220       230       240
BM10    CCTGAAG---CGCNCCCTGTGAATAACAGGTGGCTTTT--CATGGATG----TCTCTA--
        :::::::   :: .:::::::  ::   ::::: ::: :   : : :::    ::::::
PIGH    CCTGAAGGCCCGGCCCCTGTTCATCTGAGGTGCCTTCTAGAAGGAATGTGCCTCTCTAGT
              1120      1130      1140      1150      1160      1170

              250       260       270       280       290       300
BM10    GTCAGAGAAAAATGATAAAGGCTTAAATTGAGGATTAACAGAAGCAGATTAACCTCAGAA
        ::  : ::  ::   : :  :::  ::  X:::::::::::::::::::::::::: ::
PIGH    GTTGGGGAGGAACTGT-AGTGCTGCAACCGAGGATTAACAGAAGCAGATTAACCTCAAAA
              1180      1190      1200      1210      1220

              310       320       330       340       350       360
BM10    ATCCTGTCTGGCTGGCAGATTTCAAGTAA
        : ::   ::::: :::::::::::::X
PIGH    AACCCTCCTGGCCGGCAGATTTCAAGTTT
               1240      1250
```

B

```
PSBM6:      16 RTLGKGNFAVVKLARHRVTKTQVAIKIIDKTRLDSSNLEKIYREVQLMKLLNHPHIIKLY 195
               +T+GKGNFA VKLARH +T  +VAIKIIDKT+L++++L+K++REV++MK+LNHP+I+KL+
HUMP78A:    60 KTIGKGNFAKVKLARHILTGREVAIKIIDKTQLNPTSLQKLFREVRIMKILNHPNIVKLF 119

PSBM6:     196 QVMETKDMLYIVTEFAKNGEMFDYLTSNGHLS*NEARKTFWQILSAVEYCHDHHIVHRDLXTE 384
               +V+ET   LY++ E+A  G +FDYL ++G++  +EAR  F+QI+SAV+YCH+ +IVHRDL +E
HUMP78A:   120 EVIETQKTLYLIMEYASGGKVFDYLVAHGRMKEKEARSKFRQIVSAVQYCHQKRIVHRDLKAE 182
```

FIG. 2. Representative DNA and amino acid alignments of BM cDNAs. (*A*) FASTA alignment of BM10, an unknown with restricted expression, and pig unidentified hepatic protein (PIGHEP1, accession no. M29072, 72.9% identity in 177-nt overlap). The Xs delineate the area of maximal identity (indicated by pairs of dots). (*B*) BLASTX amino acid alignment of PSBM6 generated from a single-pass DNA sequence determination with human p78 (HUMP78A, accession no. M80359); $P = 6.1 \times 10^{-53}$. As shown in the middle sequence, identities = 70/123 (57%), and conservative subsitutions (denoted by +), plus identities = 105/123 (85%). The frame of PSBM6 equals +1. Numbers refer to base number except for human p78.

sion. In human BM this may be as high as 20% and may be a consequence of the variety of cell types present in total BM.

Other criteria for the success of shotgun cDNA sequencing projects are the quality of the library and the techniques used to enrich for novel sequences. This total BM cDNA library was constructed and sequenced without amplification. Success can be easily monitored by the lack of rare mRNAs being sequenced with high redundancy. The other important criteria are the technique or combination of techniques used to enrich cDNA populations for novel sequences. These include cDNA library normalization (23), subtraction (3), and differential screening (22). We employed a combination differential screen to remove abundant sequences; this technique can be extended to include multiple rounds of differential screens to remove clones that have been previously sequenced. An additional advantage of differential prescreens is the removal of redunant sequences of clones that may be overrepresented in a cDNA library due to amplification. With additional advances in cDNA library construction and screening, the technology exists for large-scale shotgun sequencing of cDNA, whereby it should be possible to identify a "gene profile" of the expressed cDNAs for a given tissue or cell. These data could then be compared to a gene profile of the tissue at a later developmental or disease state to identify changes in gene expression.

Another Approach to Development. Clearly, the analysis of large numbers of cDNAs from a particular tissue affords the opportunity to generate many markers for cellular differention. When tumors or normal cells fractionated by physiological properties (e.g., by cell sorting) from a particular tissue are tested for the expression of these cDNA markers, distinct cell lineages and individual stages within particular lineages may be defined (5). Accordingly, the sequence and expression analysis of cDNAs from complex organs such as the breast, prostate, or BM may afford a powerful new approach to studying the normal and pathological development of complex organs and organ systems.

Note. While this paper was in preparation, the sequence of PSBM43 (unknown) appeared in the data bases as an unpublished human differentiation-dependent A4 protein (HUMA4).

We gratefully acknowledge Mike Fogliano for assistance with the computer analysis. This work was supported by the Lucille P. Markey Charitable Trust (S.L.O.), the Cancer Research Institute, and the Jaye Haddad–Concern Foundation Fellowship (T.P.H.). C.L.S. is a Howard Hughes Medical Institute Research Fellow and O.N.W. is an Investigator of the Howard Hughes Medical Institute. Portions of this work were funded by grants to O.N.W. from the National Cancer Institute.

1. Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. & Venter, J. C. (1992) *Nature (London)* **355**, 632–634.
2. Caskey, C. T. & Rossiter, B. J. (1992) *J. Pharm. Pharmacol.* **44**, 198–204.
3. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. & Venter, J. C. (1991) *Science* **252**, 1651–1656.
4. Hoog, C. (1991) *Nucleic Acids Res.* **19**, 6123–6127.
5. Orr, S. L., Gese, E. & Hood, L. E. (1992) *Mol. Biol. Cell* **3**, 761–773.
6. Gieser, L. & Swaroop, A. (1992) *Genomics* **13**, 873–876.
7. Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* **162**, 156–159.
8. Koeffler, H. P. & Golde, D. W. (1978) *Science* **200**, 1153–1154.
9. Lubbert, M., Herrmann, F. & Koeffler, H. P. (1991) *Blood* **77**, 909–924.
10. Weiss, A., Wiskocil, R. L. & Stobo, J. D. (1984) *J. Immunol.* **133**, 123–128.
11. Foley, G. E., Lazarus, H., Farber, S., Uzman, B. G. & McCarthy, R. E. (1965) *Cancer* **18**, 522–526.
12. Tucker, K. A., Lilly, M. B., Heck, L. J. & Rado, T. A. (1987) *Blood* **70**, 372–378.
13. Pegoraro, L., Matera, L., Ritz, J., Levis, A., Palumbo, A. & Biagini, G. (1983) *J. Natl. Cancer Inst.* **70**, 447–453.
14. Lange, B., Valtieri, M., Santoli, D., Caracciolo, D., Mavilio, F., Gemperlein, I., Griffin, C., Emanual, B., Finan, J., Nowell, P. & Rovera, G. (1987) *Blood* **70**, 192–199.
15. Martin, P. & Papayannopoulou, T. (1982) *Science* **216**, 1233–1235.
16. Altschul, S. F., Gish, W., Miller, W., Myers, W. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
17. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
18. Uberbacher, E. C. & Mural, R. J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 11261–11265.
19. Ross, J., Gielen, J., Packman, S., Ikawa, Y. & Leder, P. (1974) *J. Mol. Biol.* **87**, 697–714.
20. Donohue, D. M., Gabrio, B. W. & Finch, C. A. (1958) *J. Clin. Invest.* **37**, 1564–1570.
21. Sibille, C., Chomez, P., Wildmann, C., Van Pel, A., De Plaen, E., Maryanski, J. L., deBergeyck, V. & Boon, T. (1990) *J. Exp. Med.* **172**, 35–45.
22. Wilcox, A. S., Kahn, A. S., Hopkins, J. A. & Sikela, J. M. (1991) *Nucleic Acids Res.* **19**, 1837–1843.
23. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. (1988) *Science* **241**, 58–62.